5264–5272 Nucleic Acids Research, 1993, Vol. 21, No. 22

MUST, a computer package of Management Utilities for Sequences and Trees

Hervé Philippe
Laboratoire de Biologie Cellulaire, URA CNRS 1134 D, Bâtiment 444, Université Paris-Sud, 91405 Orsay Cedex, France

Received April 22, 1993; Revised and Accepted September 20, 1993

ABSTRACT

The MUST package is a phylogenetically oriented set of programs for data management and display, allowing one to handle both raw data (sequences) and results (trees, number of steps, bootstrap proportions). It is complementary to the main available software for phylogenetic analysis (PHYLIP, PAUP, HENNIG86, CLUSTAL) with which it is fully compatible. The first part of MUST consists of the acquisition of new sequences, their storage, modification, and checking of sequence integrity in files of aligned sequences. In order to improve alignment, an editor function for aligned sequences offers numerous options, such as selection of subsets of sequences, display of consensus sequences, and search for similarities over small sequence fragments. For phylogenetic reconstruction, the choice of species and portions of sequences to be analyzed is easy and very rapid, permitting fast testing of numerous combinations of sequences and taxa. The resulting files can be formatted for most programs of tree construction. An interactive tree-display program recovers the output of all these programs. Finally, various modules allow an in-depth analysis of results, such as comparison of distance matrices, variation of bootstrap proportions with respect to various parameters or comparison of the number of steps per position. All presently available complete sequences of 28S rRNA are furnished aligned in the package. MUST therefore allows the management of all the operations required for phylogenetic reconstructions.

INTRODUCTION

Over the last decade, the use of nucleic acid and protein sequences in phylogenetic research has steadily increased. Sequence comparison studies have addressed the phylogeny of species at very different evolutionary scales, for example from the universal tree of life to the history of the human species (for reviews, 1–5), but also the phylogeny of genes; indeed sequence analysis permits a more natural classification of genes than the use of phenotypic characters such as enzymatic or pharmacological properties (for a classic example see 6, and for a recent one, 7).

The nature of the molecular data lends itself well to computerized analysis by the two widely used methodologies of phylogenetic analysis, phenetics (which uses genetic distance estimates between pairs of taxa) and cladistics (which uses discrete character states of nucleotides or amino acids) (8–12). In addition, the advances in sequencing methods (13–15) and the increased emphasis on molecular approaches in biology, generate considerable amounts of data which necessitate efficient computerized management.

Few programs which integrate all the data management operations necessary in phylogenetic reconstruction have been written (e.g. CLUSTAL and UWCG). Up to now, only two steps, sequence alignment and the construction of trees, have been extensively covered. Indeed, a large number of programs exist for sequence alignment (for example, 16–18) and for tree construction (PHYLIP from Felsenstein, PAUP from Swofford, HENNIG86 from Farris, MacCLADE from Maddison and Maddison).

The MUST package described in this paper manages all the operations required for phylogenetic reconstructions. It is an ensemble of programs for data management and display, allowing one to handle both raw data (sequences) and results (trees, number of steps, bootstrap proportions). It is complementary to all the major phylogenetic software cited above, and fully compatible with it; indeed, MUST is of little use without these programs. The package includes four major sections, each of them offering numerous options detailed below: (i) Acquisition, storage, modification and checking of molecular data; (ii) Sequence editing and aligning; (iii) Selection of data (taxa, sequences) for phylogenetic reconstruction; (iv) Critical analysis of the phylogenetic inferences.

MATERIALS AND METHODS

The package is written in Microsoft C 5.1 and runs exclusively on a PC using MS-DOS Version 3 or later. The software installation requires approximately 5 Mb of hard disk, to which several Mb must be added for the files created while using it. It does not work in extended memory, but requires about 550 Kb of the 640 Kb of base memory.

A VGA type or compatible screen is necessary; for some options, a color display is recommended. Major printouts (trees, matrix comparisons, or bootstrap proportion displays) are produced in graphic mode and thus require a dot matrix printer (preferably one with 24 pins) or a Postscript printer. Among dot matrix printers, only the 24 pin NEC printers have been

successfully tested for all the graphic outputs. Tree printing has also been tested successfully on numerous printers, including EPSONs and IBMs. MUST is not compatible with all existing printers; however, some printouts are in text mode and therefore compatible with all printers.

RESULTS

Flowcharts: an overview of the MUST package

The flowchart schematic of MUST (Figures 1 and 2) demonstrates how it interacts with the main programs of phylogenetic reconstruction (PHYLIP, PAUP, CLUSTAL, and HENNIG86). MUST has been conceived for the best possible management of a large amount of sequences for phylogenetic purposes, from sequence acquisition to tree construction. Towards this end, the storage of sequences is done in a phylogenetic framework from the start. Thus, during the storage procedure, the sequence is associated with the name of the species from which it was obtained; this name is itself associated with a higher taxon name. Consequently, species are located within a cladogram, defined by the user (MAKETREE). As a result, the user is confronted by a phylogenetic question in every taxonomic choice (in the course of the various programs).

Various programs store new sequences in the database while associating them with numerous comments, one being the group name of the species, which allows one to locate it in the predefined cladogram (systematic frame, Figure 1). These new sequences may be extracted from databanks (EMBL, GENBANK, NBRF and SWISS-PROT) or may be entered by the user (ENTRYSEQ). The sequences are protected from all unintentional errors, but MODIFSEQ allows their modification, while displaying explicitly the modifications performed. The functions of a data bank, such as deleting, displaying, or saving data, are fulfilled by the programs DELBANK, INFOBANK, SAVEBANK, and LOADBANK.

The link between the sequence bank and the files of aligned sequences is achieved either when storing them in the bank for the first time (SAVESEQ, EMBL, etc.) or afterwards (MODIFSEQ, INTEGSEQ, and INTEGALL). VERIFSEQ checks that the sequences contained in a file of aligned sequences have not been unintentionally modified, and are maintained identical to the bank version.

The alignment determines the homologies in the sequences. The programs of automatic alignment are generally well equipped for the bulk of this work, but it is most often necessary to adjust

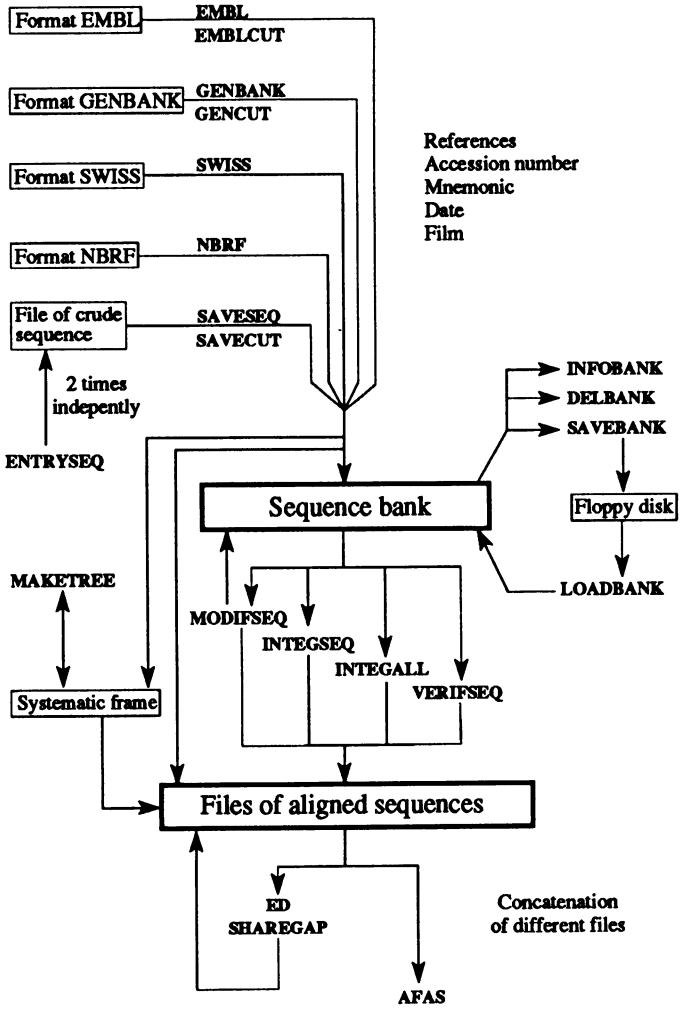

Figure 1. Schematic flowchart of the management of raw sequences. The names of programs are in bold characters. The different types of files are written in a frame and the arrows show the information transfers.

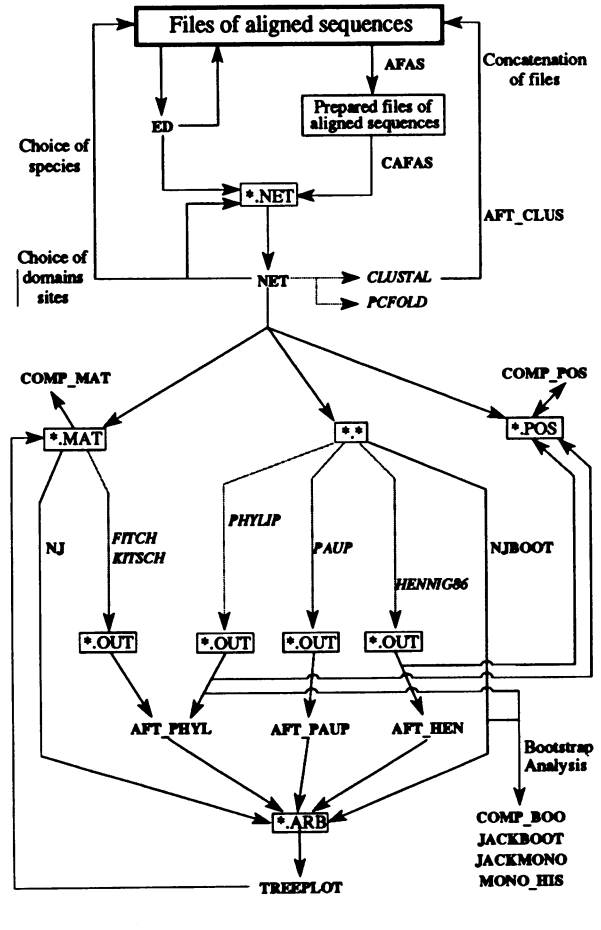

Figure 2. Schematic flowchart of the management of aligned sequences. The names of programs are in bold characters when they have been written by the author, and in italic when not. The different types of files are written in a frame and the arrows show the information transfers.

the alignment by hand. This is why MUST produces and reads files used by CLUSTAL and contains an editor of aligned sequences (ED) with numerous functions.

The entire dataset available is rarely used to construct a phylogeny. MUST makes selection of species easy using the predefined cladogram (CAFAS and ED). A selection of sites is especially necessary to eliminate poorly aligned characters for which homology cannot be clearly determined; it also permits removal of the non-informative characters which slow down the reconstruction programs. NET offers the possibility of eliminating regions, either directly, from the display of the alignment, or by using indices (e.g. number of characters present at a site or position in the codon).

N*Et also* formats files for numerous phylogenetic reconstruction programs (PAUP, PHYLIP, and HENNIG86). AFT_PAUP, AFT_PHYL, and AFT_HEN transform the outputs of these programs to make them usable by other programs of the package. NJ (based upon the Neighbor Joining method of Saitou and Nei (19) and NJBOOT (applying the bootstrap procedure (20) on the Neighbor Joining method) are the only two reconstruction programs provided in MUST.

After having constructed a phylogeny, one must be able to display it and especially to analyze critically the inferences drawn from the phylogeny. TREEPLOT displays and prints the tree constructed by each of the programs presented in Figure 2. It allows one to choose the position of the root and to rotate the branches around the nodes. A critical analysis of results is possible by comparing matrices (COMP_MAT), bootstrap proportions (JACKBOOT, COMP_BOO, JACKMONO, and MONO_HIS) and number of steps per position (COMP_POS).

Management of non-aligned sequences

The purpose of this part of the package is to create a database adapted to the problems with which the molecular phylogeneticist is faced. This management prevents unintentional alterations of sequences in the computer (non-Darwinian evolution!), and identifies sequences by specific comments.

In the course of the various programs, each sequence is recognized by three keywords: taxon name, molecule and domain of the molecule. The first keyword which is the identifier of the organism sequenced follows the format: (genus name) (species name)_(various characteristics). It can contain up to 40 characters, and will be used unmodified throughout all the steps of the phylogenetic reconstruction, up to the final tree.

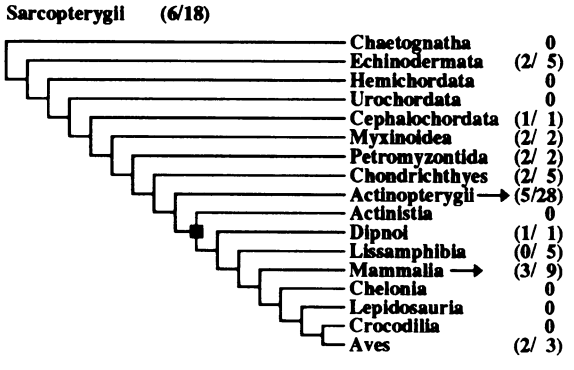

Figure 3. Example of a cladogram from which one chooses species. The cursor is located on an internal node, the Sarcopterygii, displaying a total of 18 species out of which 6 species have been selected. For the terminal nodes, the column on the right indicates the number of available species in the database and the column on the left indicates the number of currently selected species. Detailed cladograms of the groups outlined by an arrow can be displayed on a subsequent screen.

The two other keywords associate the sequence with a type of molecule. As a general rule, the name of the gene suffices to describe a sequence. However, it can be useful to cut large genes into different domains of homogeneous variability (see below). In this case two keywords are needed to identify the sequence: Molecule and Domain.

The program ENTRYSEQ allows the acquisition of new sequences. It requires two independent entries of the sequence. When acquisition is achieved, the two sequences are displayed one below the other, in order to correct rapidly all differences. This procedure greatly reduces the probability of errors in reading or in entry.

Storing is achieved either starting from files in the format of the databanks (EMBL, GENBANK, SWISS-PROT and NBRF) or from a file containing only the sequence under study, created using ENTRYSEQ or any other program. The group name of the new sequence, which will be used to choose the species in the other programs, is given at this step.

When the sequence is imported from a databank, the user may choose the comments which will be retained, and add his own comments. The parameters 'Accession number'and 'Mnemonic', as well as the date of storage, are automatically stored. The user may choose the portion of sequence that will be stored, either by using the Feature Tables, or by giving explicitly the boundary values. For example, one may select the 100 nucleotides preceding the ATG start, or all the exons of a gene. After selection, the user can invert the sequence portion, obtain the complement, or translate the nucleotide sequence into amino acid sequence.

In the case of a de novo storage, all available information concerning the organism and the sequencing must be kept. This is why, in addition to the field of free comments, the following predefined fields exist: Characteristics of the organism, Author and quality of the sequence, Number and reader of films.

After storage in the databank, the new sequence may be incorporated into an already existing or a new file of aligned sequences. This incorporation may also be performed later (INTEGSEQ and INTEGALL). When the user modifies the sequences, MODIFSEQ displays the previous version and indicates in red all the modifications made in order to minimize errors. Finally, MUST also provides the general functions of a database (deletion of elements of the base with DELBANK, storage of the base with SAVEBANK, restoration with LOADBANK and information searches with INFOBANK).

The principal advantage of such a bank lies in the possibility to check, at any given time, with the program VERIFSEQ, the agreement between the sequences of a file of aligned sequences and those stored.

Choice of species

The choice of species is of primary importance in improving alignment and constructing a phylogeny. In MUST, the species are chosen from a cladogram (Figure 3). The cursor is moved over all the nodes of the cladogram; at the same time, the number of species of the group corresponding to the active node is displayed. A function allows selection of all the species of the active group, but a partial selection, species by species, is also possible. A general cladogram of living organisms is provided,

but the user may construct a cladogram corresponding more precisely to his needs, for example, a cladogram of the Eubacteria or a cladogram of the Artiodactyla.

Display modes of the sequence editor

In order to make an alignment de novo or optimize an existing alignment, MUST includes a sequence editor, ED. The user chooses one or several subgroupings of species for which he wishes to make or modify an alignment. The gaps common to all the species of the subgrouping are not displayed on the screen, which much improves the visual impression and is particularly valuable when scores of species are handled. Nevertheless, the information about such gaps remains available. The absolute position of the cursor in the sequences (i.e. taking the hidden gaps into account) is displayed: the user knows that gaps occur between two positions if the position displayed is increased by more than one while the cursor is moving one position further on the screen.

Moreover, the display of aligned sequences can be adjusted to the user's convenience: (i) replacement of characters identical to those of the upper species on the screen with a dash, or not; (ii) blocks of 10 or of 60 characters; (iii) use, or not, of a specific background color for each type of character. In this last option, the user can choose, for each nucleotide or amino acid, the background color to be used. It is thus possible to visualize the conservative changes in the proteins by assigning the same color for one amino acid type: hydrophobic, acidic, basic...

Finally, the choice of the upper species on the screen conditions the perception of the alignment, when one visualizes the aligned sequences while replacing the characters identical to those of the upper species by a dash. The species on the first line can be easily exchanged with any of the others under analysis to avoid alignment errors due to this effect.

Modification of an alignment

In a file of aligned sequences, the user can only add blanks and stars in the sequences and cannot modify the sequences themselves. The stars indicate an insertion or a deletion, whereas the blanks (space bar) indicate the absence of a part of the sequence. The user can add blanks and stars one by one, or by blocks of any size. Similarly, one can delete these characters one by one or by continuous blocks: in this case, ED suppresses all the stars and blanks located on the right of the cursor up to the first nucleotide or amino acid. Adding or deleting stars by the block noticeably accelerates the alignment.

When only one species contains an insertion, ED allows the addition, in a single operation, of one or several stars in all the other sequences in the file (including those which have not been selected) but not in the sequences on which the cursor rests. This function prevents undesired modifications of the former alignment when a species is added.

Search for local similarity

The large biological molecules are mosaics of domains evolving at different rates: divergent versus conserved domains of ribosomal RNAs, introns versus exons, transmembrane segments versus cytosolic loops. A zone where alignment is difficult if not impossible can therefore occur at 50 or 100 nucleotides to the right of a highly conserved block where alignment is straightforward. The visual search for the next conserved block is difficult even when the similarity is very strong. A function of ED allows a rapid search for similarity between the upper sequence on the screen and a sub-sequence. The first 20 characters on the right of the cursor having been automatically taken as a reference, the upper sequence is scanned to determine the best similarities. The fifteen strongest similarities are displayed in decreasing order (Table 1). For each of the fifteen regions, the absolute position of the beginning of the block and its relative position with respect to the cursor are indicated. The reference sub-sequence may also be acquired manually. By this function, the user knows immediately whether the current region can be aligned with the upper sequence on the screen, and how much displacement is necessary to obtain that alignment.

Table 1. An example of the 15 strongest similarities between a selected sub-sequence of 20 nucleotides and a complete 18S rRNA as displayed by the ED program. The computer has found, almost instantaneously, a very likely homologous portion far from the current position of the cursor.

Number of matches in a total of 20	Positions in the sequence	Distance to cursor
16	496	137
11	1229	870
10	612	253
10	703	344
10	1218	859
10	1433	1074
10	1977	1618
9	15	-344
9	46	-313
9	97	-262
9	410	51
9	418	59
9	468	109
9	502	143
9	517	158

Consensus sequences

The user can create different subgroupings of species (disjointed or not) and display them simultaneously. For a given position, ED seeks the most frequent character state in a subgrouping and symbolizes the frequency of occurrence of that character by a specific foreground color. The consensus sequence so determined illustrates the variability of the sequences in the subgrouping. For a better visual perception, the brightness of the color diminishes with the frequency of the most frequent character state.

The consensus sequence is managed like a species sequence. It is thus possible to add or delete stars within the consensus sequence. In this way, the user modifies the alignment of a subgrouping with respect to others without modifying the alignment inside it. This function allows one to align the subgroupings by putting the emphasis on regions of strong consensus, after alignment has been performed inside each one. This method greatly accelerates the obtaining of alignment and allows the management of large files of aligned sequences containing several hundred species.

Optimization of alignment

It is necessary to check whether stars have been added without improving the quality of alignment. One often creates unintentionally cases such as following:

```
Subgroup 1 ATGATAG*ATGGACGAT*TTTGGTCGATGACGAT
Subgroup 2 ATGGTAG*ATTGACGATTTT * GGTTGATGAAGAT
                                 ↑
```

The program SHAREGAP seeks all the positions where more than 50% of the species have a star and then displays the corresponding regions (30 characters on each side of the position with more than 50% gaps) as two consensus sequences, one of the species sharing the gap and the other grouping the remaining species.

The user may then rectify the position of the star with ED if necessary. In addition, the program SHAREGAP allows, for a specified file, the search for and elimination of positions where all the species have a star or a blank, positions which are often created when one modifies the alignment.

Concatenation of files of aligned sequences

The obtaining of reliable phylogenetic trees generally requires a very large amount of molecular information (20–22). Usually a single gene or portion of gene does not provide enough information. It often proves necessary to use several (23–25), and thus to concatenate sequences from different genes.

The program AFAS (Addition of Files of Aligned Sequences) accomplishes this task very rapidly by either of two options: keeping only the species which have all the regions (intersection), or keeping all the species present in at least one of the regions while adding blanks when the region does not exist (union). Owing to the rapidity of concatenation, the user may obtain phylogenies with different genes taken separately as well as altogether (partial versus total evidence).

Segmentation of large genes

The large genes (18S and 28S rRNA, elongation factors, polymerases) are composed of contiguous domains evolving with very different substitution rates. As a result, some portions are similar in distantly related species, whereas others may be aligned in closely related species only. Visualization becomes difficult and alignment inefficient when these two types of regions are included in a single file. For example, the alignment of the D2 domain of 28S rRNA (26) from Archebacteria (30 nucleotides) necessitates more than 700 completely useless stars if mammalian species are also present (approximately 800 nucleotides). From a practical point of view, splitting large genes into domains of similar variability is useful. MUST manages such cuts efficiently, by the concatenation program described above, and programs that allow the splitting of a gene. The latter (SAVECUT, EMBLCUT, GENCUT) work like the other storage programs (SAVESEQ or EMBL), but in addition seek sub-sequences which serves as limits of conserved domains and split the gene according to these limits. This splitting avoids the addition of hundreds of stars owing to species which are distantly related and which, in any case, will not align in variable regions.

Extraction of a subgrouping of species

The extraction of species from the files of aligned sequences is achieved in the same manner in the programs ED and CAFAS (Choice in Adding Files of Aligned Sequences). Species are selected from a predefined cladogram as explained above. Since the user often needs to use species lists which are nearly the same, any list of species may be saved in a file. When he loads it later (whether for the same file of aligned sequences or not) this will avoid having to select the species anew and he can rapidly make small modifications (additions or eliminations). Once the species are chosen the selected sequences are written into a file which contains all the information necessary for further analysis.

Table 2. The different ways for distance calculations in MUST.

DNA or RNA
All the differences
Transitions
Transversions
Deletions
Transitions + Transversions
Transitions + Deletions
Transversions + Deletions
Distance of Jukes and Cantor
Distances of Kimura (2 parameters)
PROTEINS
Boolean (any difference = 1)
Miyata (hydrophobicity and encumbrance)
Hydropathy index (Kyte and Doolittle)
Similarity of structures II (Risler)
Log-odds matrix (Dayhoff)

Extraction of a subset of positions

While the selected aligned sequences are displayed, the user may 'mark' the regions to be excluded in further steps. It is imperative to use this option to eliminate the regions where the alignment is dubious, because the corresponding characters are not homologous and should not be used in phylogenetic analyses. One may also use this option to split a gene into two portions and compare the results obtained independently from both (see 27). The program NET can also eliminate all the positions which contain more than a certain percentage of a given type of character; this applies to the character states indicating (i) the absence of sequence (blank), (ii) a gap (*), or (iii) an imperfectly determined nucleotide (R, Y, M, W, S, K, D, H, V, B, or N) or amino acid. The number of sites eliminated by retaining different percentage values (0, 1, 2, 3, 4, 5, 10, 25, 50, and 99%) of such states as a criterion of elimination of the sites are displayed. The user then chooses from that list the limit percentage value. Thus, in three operations, one can eliminate all the sites containing a blank, a star, or an imperfectly determined nucleotide.

The choice of poorly aligned regions for elimination is a long and delicate (because subjective) task. It is thus preferable to do it only once, carefully, then to make all the analyses with that selection. MUST allows one to proceed in this manner by storing and retrieving the boundary values of eliminated regions. Furthermore, MUST always uses the absolute positions in the starting files of aligned sequences, even if, for a subgrouping of species, the positions contain only stars and blanks (common deletions) and thus disappear on the display. The boundary values are therefore the same in two different subgroupings of species (as long as they come from the same file of aligned sequences) despite the different common deletions. This function allows one to work on the same set of homologous characters in different sets of species.

It is also possible to discard all the sites corresponding to a given criterion simultaneously: number of character states present at a position, number of character states present at least two times

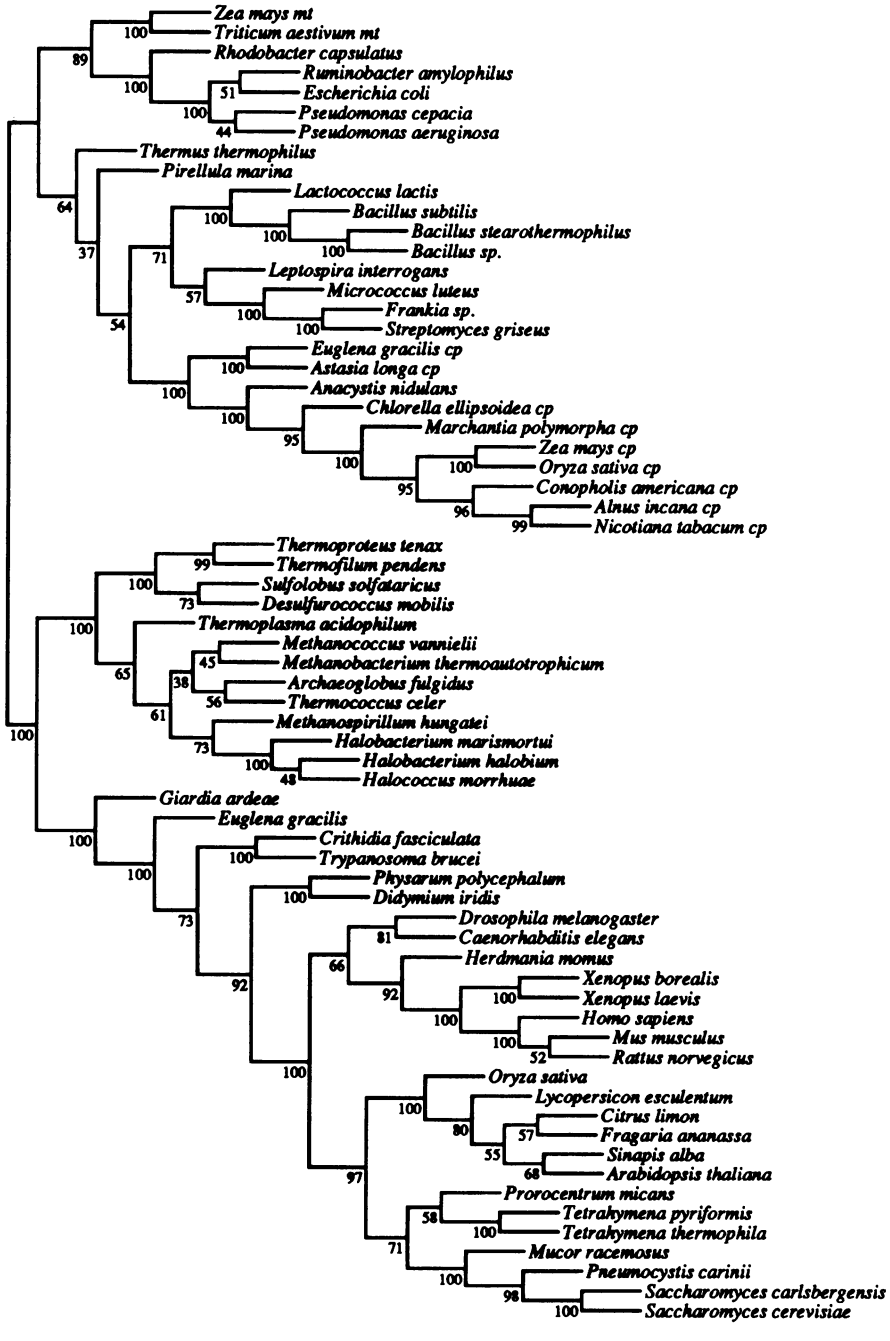

Figure 4. Example of a tree, constructed by NJBOOT and printed by TREEPLOT, for all available species (67) for the complete 28S rRNAs (provided along with MUST). The NET program has allowed the exclusion of the more variable regions, the positions containing at least one gap and the non-informative positions; this leads to a total of 962 informative sites. The bootstrap proportions are indicated for each node, in percentages.

at a position, and position in the codon. One may easily eliminate constant sites, sites uninformative for parsimony, sites where more than 5 different character states occur, the third base of codons, or the first two bases of codons.

Data formatting

After the previously described operations have been performed, sequences or distance matrices can be formatted by MUST for numerous programs, principally those for phylogenetic reconstruction: NJ and NJBOOT (MUST), PAUP (Swofford), PHYLIP (Felsenstein), HENNIG86 (Farris) and MACCLADE (Maddison and Maddison) for the construction of trees, and CLUSTAL (Higgins), PCFOLD (Zucker), NET and ED (MUST) otherwise. In addition, a variety of procedures for calculating genetic distances are included in MUST (Table 2).

For the parsimony programs, three types of coding are proposed. The first leaves the sequences unmodified in order to use all changes (substitutions and gaps). The second replaces the stars by a question mark, in order to use only the substitutions. In the case of nucleic acids, the third replaces the stars by a question mark, the purines by a G, and the pyrimidines by a C in order to use only the transversions.

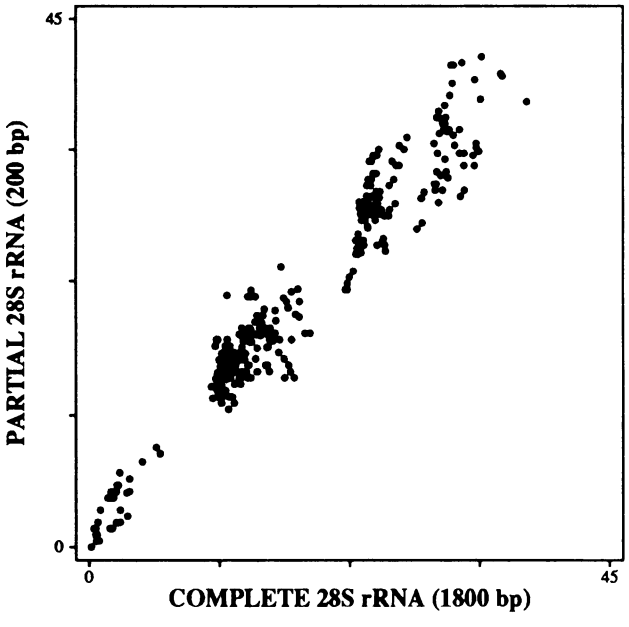

Figure 5. Comparison of matrices for 26 eukaryotes printed with COMP_MAT. On the X axis, values correspond to distances calculated with the 1800 nucleotides of the complete aligned 28S rRNA, while values on the Y axis correspond to distances calculated with only 200 nucleotides of 28S rRNA. A correlation coefficient of 0.97 is obtained, demonstrating that few nucleotides are needed to obtain a reliable estimate of the distance between two species.

Any new output format, method of calculation of distances, or coding of characters may be added on request.

Construction of phylogenetic trees

Only two programs of tree construction are included in MUST: NJ (Neighbor Joining) and NJBOOT. NJBOOT creates three output files: one containing the majority rule consensus tree, another the most frequent groups of species, and the last the groups occurring in more than 1% of the replicates. These programs are rapid. For instance, on a PC 486 at 50 MHz, NJBOOT treats 67 species and 962 characters in 77 minutes for 1000 replicates (tree of Figure 4).

These two programs are not sufficient for a phylogenetic analysis. They are complemented by the other programs for which MUST has a formatted output (HENNIG86, PAUP and PHYLIP). The outputs of these programs (topology, plus branch lengths and number of steps per position when possible) are automatically reformatted in order to be utilized by MUST.

Display and printing of trees

TREEPLOT displays the trees previously stored in parenthesized form. The user may, in an interactive manner, position the root anywhere, pivot the two daughter branches around each node, and align the species at the right. The printout is the exact reproduction of the screen display. Thickness of the lines, space between species, and scale may be adjusted. The printing is implemented for printers at 8, 9, and 24 pins, and for Postscript printers.

Intermediate files summarize all previous treatments

MUST creates numerous intermediate files allowing one to stop and restart at any step of the phylogenetic study. These files contain all the previous information: starting file of aligned sequences, date of creation, and name of program used, etc. For example, one tree file may contain the following comments:

```
# Sequences extracted from 18SAA.ALI of the 20 July 1992 at
  15 hours 4
# File 18S.NET created on Friday 11 December 1992 at 13 hours
  32
# File 18S_0.NBS created on Friday 11 December 1992 at 13
  hours 35
# Line used for free comments
# Type of parsimony: All the transversions
# BOUNDARIES: 200-443 635-698 804-1041 1390-1443
  1710-1804 2128-2264
# 1022 sites eliminated because PRESENT2 ∈ [2.00, 5.00] (∈
  [1.00, 5.00]
# 328 positions remain on the 1350 selected (total=2114)
# Tree calculated by the program HENNIG86
# mhennig length 421 ci 49 ri 58 trees 5
```

Table 3. The maximum number of species and characters, and the maximum value of the product of these two parameters, for the most limited programs.

Program	Number of species	Number of characters	Sp * Char
ED	400	25000	200000
AFAS	300	20000	200000
CAFAS	1000	20000	
NET	300	10000	150000
NJ	180		
NJBOOT	100	32000	65000
COMP_MAT	140		

Critical analysis of data and inferences

Several tools for evaluating the robustness of the inferences or the congruence of two data sets have been incorporated in the package. The comparison of distance matrices is one of these tools. The idea is to check whether two independent sets of genetic distances obtained among a number of species are congruent. The two sets can correspond to two different genes (such as 18S and 28S rRNA or two protein coding genes, etc.), or two different portions of the same gene, or two different ways of calculating the distances in the same dataset, etc. COMP_MAT compares the distances contained in two files. Species common to the two files are first sought; then all the possible pairs of species are displayed in an X-Y plot (Figure 5). Each pair of species has, on the abscissa, its distance in the first file, and, on the ordinate, its distance in the second file. This comparison can make saturation of transitions with respect to transversions obvious (28); it can also show saturation in substitutions of the whole gene by comparing the patristic distance (calculated using an option of the program TREEPLOT after the calculation of the branch lengths by PAUP) and the phenetic distance (29). In addition, the pair of species corresponding to each point may be identified, which is particularly useful for discussion of extreme points (for example, due to an acceleration of the evolutionary rate of a species for a gene). Finally, the points corresponding to all the pairs of species including any given species may be displayed; in this way one may discover, for example, for which species the gene under study is saturated.

The bootstrap proportion (BP) is now a frequently used index of the statistical validity of a node. We have incorporated a set of programs allowing the calculation of BP as a function of a number of parameters. The variation of BP may be displayed

as a function of two parameters: sequence length and choice of species. In the first resampling method, the sites or species are sampled without replacement and various programs compare the BP obtained with a subgrouping to the BP obtained for the whole data set. In the second one, all the combinations of single species per predefined monophyletic groups are submitted to bootstrap analysis. These different programs (COMP_BOO, JACKBOOT, JACKMONO, and MONO_HIS) and their use on a phylogeny of Gnathostomes are presented in Lecointre *et al.* (22, 30). These tools permit a rough inference of the number of additional nucleotides needed for the resolution of a given node, and the precise analysis of the impact of selected species.

Finally, it is simple to study the number of steps per position. In fact, as indicated above, MUST retains the absolute position of each site in the file of aligned sequences, whatever the set of species used and despite deletion of some regions of sequences. Since each site remains identifiable whatever the file may be, MUST generates files containing either the number of characters present at a site, the number of characters present at least two times at a site, or the number of steps calculated by a program of parsimony (HENNIG86 and PHYLIP). The program COMP_POS displays an X-Y plot of the values contained in two files. In this way, one can compare the number of characters at a site with the number of steps in a tree, or the number of steps in the most parsimonious tree with the number of steps in a tree constructed with other characters. As a result, homoplasy may be critically analyzed (31).

DISCUSSION

The computer package described in this paper offers a number of advantages for managing and treating data in a phylogenetic reconstruction perspective; it also has a few drawbacks. These will be briefly discussed. Using a cladogram to help in selection of species both defines the phylogenetic problem under study and provides the current resolution of that problem, before one even begins the construction of molecular phylogenies. It allows one to formulate clearly the phylogenetic problem and to evaluate the quality of the sequence sampling in view of current knowledge, an important point, as bad sampling can have dramatic consequences (30). However, this procedure leads the user to try to align the species of the same predefined group, even if this group is not monophyletic. To avoid this artefact, it is necessary to be very careful about which regions are eliminated as non-alignable.

MUST very noticeably accelerates the establishment of phylogenies starting from molecular data, and avoids chance introduction of errors in the sequences. Numerous combinations (choice of species, choice of sites, choice of reconstruction programs) are easily generated and numerous parameters of phylogenetic analyses finally become testable. The critical analysis of these variations permits an empirical evaluation of the reliability of trees.

However, MUST is limited as to the number of species and characters by DOS (Table 3) since it uses only the 640 Kb of the base memory. To overcome this constraint, a UNIX version is being designed. In addition, MUST generates a very large number of files which quickly fill several Mb on the hard drive. It is advisable to plan for a minimum of 10 Mb for the files created by MUST, beyond the 5Mb needed for the installation of the package, and to delete regularly files no longer useful.

MUST is furnished with several files of aligned sequences (LSU rRNA, superoxide dismutase). In particular, all the presently available sequences of 28S ribosomal RNA, aligned in the conserved domains and non-aligned in the divergent domains, are provided. The user can therefore become familiar with the different functions, using a very rich data set, before treating his own data.

Distribution of the software

MUST is a shareware package. To obtain MUST, one may (i) write to the author enclosing four 3.5 in high density diskettes; (ii) copy the diskettes of any other person who already has the software; or (iii) send $100.00 to the author, who will then send the four diskettes along with a detailed user's manual. In the last case new versions will be provided, in particular debugging corrections; and one may ask that new functions, especially new output formats, be developed.

ACKNOWLEDGEMENTS

I especially thank A. Adoutte for having encouraged me to do this work, and B. Philippe for helping me to conceive the user interface, for participating in the editing of the documentation, and for writing some modules. I thank G. Lecointre for participation and phylogenetic information. For suggestions and informed criticisms after usage of MUST, I thank: G. Balavoine, G. Coffe, E. Douzery, V. Lê, M. Muller, and S. Tillier. I thank H. Le Guyader for his critical reading of the paper. C. Thompson-Coffe is gratefully acknowledged for her contribution to the preparation of the English version of the manuscript and of the user's manual as well as Maximilian Telford for help in polishing the final English version. This work has been supported by a BDI (Bourse de Docteur-Ingénieur) from the CNRS. I acknowledge the attribution of several grants from the 'Direction de la Recherche et des Etudes Doctorales' of the 'Ministère de l'Education Nationale' which allowed the acquisition of all the equipment used in this study.

REFERENCES

1. Woese,C.R. (1987) *Microbiol. Rev.*, **51**, 221–271.
2. Sidow,A. and Bowman,B.H. (1991) *Curr. Opin. Genet. Develop.*, **1**, 451–456.
3. Sogin,M.L. (1991) *Curr. Opin. Genet. Develop.*, **1**, 457–463.
4. Adoutte,A. and Philippe,H. (1993) In Pichon,Y. (ed.), *Comparative Molecular Neurobiology*. Birkhaüser Verlag, Basel, pp. 1–30.
5. Cann,R.L. (1991) In Warren, L. and Koprowski, H. (eds.), *New perspectives on evolution*. John Wiley and Sons, New York, pp. 209–233.
6. Goodman,M., Weiss,M.L., and Czelusniak,J. (1982) *Syst. Zool.*, **31**, 376–399.
7. Vernier,P., Philippe,H., Samama,P., and Mallet,J. (1993) In Pichon,Y. (ed.), *Comparative Molecular Neurobiology*. Birkhaüser Verlag, Basel, pp. 297–336.
8. Felsenstein,J. (1982) *Quart. Rev. Biol.*, **57**, 379–384.
9. Nei,M. (1987) *Molecular evolutionary genetics*. Columbia University Press, New York.
10. Felsenstein,J. (1988) *Annu. Rev. Genet.*, **22**, 521–565.
11. Swofford,D.L., and Olsen,G.J. (1990) In Hillis,D.M. and Moritz,C. (eds.), *Molecular systematics*. Sunderland, Massachussets, pp. 411–501.
12. Li,W.H. and Graur,D. (1991) *Fundamentals of molecular evolution*. Sinauer Associates, Inc., Publishers, Sunderland, MA, USA.
13. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
14. Qu,L.H., Michot,B. and Bachellerie,J.P. (1983) *Nucleic Acids Res.*, **11**, 5903–5920.

15. Mullis, K. and Faloona F. (1987) in Wu R. (ed.), *Methods in Enzymology*. Academic Press, New York and London, Vol. 155, pp. 335–350.
16. Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *Comput. Appl. Biosci.*, **8**, 189–191.
17. Feng,D.F. and Doolittle,R.F. (1990) in Doolittle,R.F. (ed.), *Methods in Enzymology*, Academic Press, Inc., Vol. 183, pp. 375–387.
18. Hein,J. (1990) in Doolittle,R.F. (ed.), *Methods in Enzymology*. Academic Press, inc., Vol. 183, pp. 626–645.
19. Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
20. Felsentein,J. (1985) *Evolution*, **39**, 783–791
21. Li,W.H., and Gouy,M. (1990) in Doolittle,R.F. (ed.), *Methods in Enzymology*., Academic Press, Inc., Vol. 183, pp. 645–659
22. Lecointre,G., Philippe,H., LΩ,H.L.V. and Le Guyader,H. (1993) *Mol. Phylo. Evol.* (in press).
23. Gouy,M. and Li,W.-H. (1989) *Mol. Biol. Evol.*, **6**, 109–122.
24. Graur,D., Hide,W.A. and Li,W.H. (1991) *Nature*, **351**, 649–652.
25. Bailey,W.J., Hayasaka,K., Shinner,C.G., Kehoe,S., Sieu,L.C., Slightom,J.L., and Goodman,M. (1992) *Mol. Phylo. Evol.*, **1**, 97–135.
26. Hassouna,N., Michot,B. and Bachellerie,J.P. (1984) *Nucleic Acid Res.*, **12**,3563–3583
27. Patterson,C. (1989) In Fernholm,B., Bremer,K. and Jörnvall,H., (eds.), *The hierarchy of life. Molecules and morphology in phylogenetic analysis. Excerpta Medica*, Amsterdam, pp. 471–488.
28. Leclerc,M.C., Guého,E. and Philippe,H. (unpublished).
29. Philippe,H., Sörhannus,U., Baroin,A., Perasso,R., Gasse,F., and Adoutte,A. (1993) *J. Evol. Biol.* (in press).
30. Lecointre,G., Philippe,H., LΩ,H.L.V. and Le Guyader,H. (1993) *Mol. Phylo. Evol.* (in press).
31. Philippe,H., Lecointre,G., LΩ,H.L.V. and Le Guyader,H. (unpublished).